Volume 15 Number 2 1987 Nucleic Acids Research

Sequence distributions associated with DNA curvature are found upstream of strong *E. coli* promoters

R.Richard Plaskon and Roger M.Wartell

Schools of Applied Biology and Physics, Georgia Institute of Technology, Atlanta, GA 30332, USA

Received September 9, 1986; Revised and Accepted December 12, 1986

ABSTRACT

The regions upstream from forty-three procaryotic promoters were examined for nucleotide distributions which have been associated with DNA curvature. The analysis procedure assigned a DNA curvature score based on the phasing of the 5' and 3' ends of A_n and T_n tracts, $n \geq 3$. The weighting scheme for the curvature score was based on recent studies which showed that tracts of A_n and T_n periodically phased with the helix repeat cause DNA curvature. Results show that promoters which have high transcription initiation rates in vivo tend to have high curvature scores in their upstream regions. Regions downstream from the transcription startpoint do not have sequences correlated with DNA curvature. Four promoters which have been shown to have upstream activation regions have curvature scores above 1.5 in their -40 to -150 regions. The correlations observed lend support to the hypothesis that DNA curvature is associated with upstream activation of transcription.

INTRODUCTION

Most DNA promoters of the RNA polymerase of Escherichia coli contain two highly conserved hexanucleotide sequences located about ten (-10 site) and thirty five (-35 site) base pairs behind the transcription startpoint (1). Both genetic and biochemical studies indicate that these two regions and the distance between them are of prime importance in defining the overall strength or transcription initiation rate of a promoter. In certain promoters, the deletion and/or insertion of DNA sequences 10-100 bp upstream from the -35 site also influences transcription. Changes in upstream regions decreased transcription from 2 to 15 fold (2-6). Horn and Wells (2) presented evidence for an upstream activation region in the λ P_L promoter. Lamond and Travers (3), Bossi and Smith (4), and Gourse et al. (5) have also demonstrated this phenomenon for two tRNA promoters (tyrT, hisR), and one rRNA promoter (rrnB P1). In addition, Banner et al. (6) have observed a similar region in a promoter for a Bacillus subtilis RNA polymerase.

In the studies by Bossi and Smith, and Gourse et al., the DNA regions responsible for the enhancement of transcription exhibited lower than expected

electrophoretic mobility in polyacrylamide gels. Physicochemical evidence indicates that abnormally low electrophoretic mobilities of DNA fragments can be due to a curved or bent DNA structure (7,8,9). These results raise the possibility that curved or bent DNA structures are characteristic features of upstream activation regions. As an initial step toward addressing this question, we have examined the upstream regions of forty-three promoters for distributions of nucleotides which have been associated with curved DNA structures. Trifonov (10) has suggested that curvature of the DNA helix can be generated by a distribution of ApA dinucleotides which are periodically phased with the helix repeat. Hagerman (7,11), Diekmann (8), and Koo et al. (9) have shown that stretches of A_n and T_n ($n > 3$), periodically phased, are the major cause for the intrinsic curvature indicated by electrophoresis experiments. Searches were made for both types of distributions. A positional correlation analysis was employed to screen for periodic positioning of ApA and TpT dinucleotides. An empirically based set of rules was formulated to determine a 'DNA curvature score' from the locations of A_n and T_n stretches.

The results show that promoters which have high transcription rates tend to have high curvature scores in their corresponding upstream regions. Seventeen out of forty-three promoters examined, including the promoters known to be influenced by upstream sequences, have DNA curvature scores between 12.3 and 1.55 in the -40 to -150 region. Weak promoters have curvature scores of zero to 1.2 in their upstream regions. In sequences downstream from the promoters, +60 to +170, available in thirty-six out of the forty-three promoters examined, only three regions have curvature scores above 1.2, the highest value being 3.1. This suggests that DNA curvature may play a role in enhancing transcription. The periodic positioning of ApA dinucleotides is also observed for several upstream promoter regions. However, several high transcription rate promoters including ones with known upstream activation regions do not show periodic phasing of ApA sequences. Mechanisms by which DNA curvature may influence promoter strength are discussed.

METHODS

DNA Curvature Score

An algorithm was developed which assigns a DNA sequence a numerical score based on the phasing of tracts of A_n and T_n. The rules of the algorithm are based on the gel mobility studies of Koo et al. (9), Diekmann (8) and Hagerman (7). We have employed a simple weighting scheme which attempts to reflect the

relative curvature of various DNA sequences determined from the gel mobility measurements. Future experiments using additional DNA sequences and standardized conditions will undoubtedly refine this approach.

The computer algorithm first inputs a given 110 bp. upstream or downstream region. The DNA sequence is first scanned for tracts of A_n and T_n with $n > 3$. The base pair locations of the 5' and 3' ends of these tracts are determined. Weights are then assigned for each occurence in which the 5' end of an A or T tract is separated from the 5' end of another A or T tract by a distance close to 10.5 bp. A weight of 1.5 is assigned when the distance, d, is 10 or 11 bp. The weight is 1.0 if d is 9 or 12 bp. When the distance between two 5' ends is $20 \leq d \leq 22$ bp, a weight of 0.75 is given. A similar procedure is employed for the phasing of the 3' ends of A or T tracts. The weights are 1.6 when d is 10 or 11 bp, 1.1 for d equal to 9 or 12 bp, and 0.8 for $20 \leq d \leq 22$ bp. A and T tracts which are three bases long are treated differently. A weight of 0.6 is assigned when a 3' or 5' end of one of these tracts is a distance of 10 or 11 bp from a corresponding 3' or 5' end of an A or T tract ≥ 3 bp long. We have not differentiated between A and T tracts since studies have shown that the periodic phasing of mixtures of these two tracts produces intrinsic bending (8,9). Although the length of a tract between 4-9 bp has an effect on curvature (9), we have omitted this information in the interest of simplicity. Hagerman (11) has shown that periodic phasing of A_jT_j tracts, $j = 3,4$ produces DNA curvature while T_jA_j tracts do not. We have therefore scanned sequences for A_3T_3 tracts designating their location by the position of the A/T junction. If an A_3T_3 tract is separated from a similar tract by d equal to 10 or 11 bp a weight of 1.5 is assigned. If d equals 9 or 12 the weight is 0.6, and if $20 \leq d \leq 22$ bp the weight is 0.6. The score for a given DNA sequence is obtained by adding up all of the weights.

Positional Correlation Analysis

This analysis is similar to that described by Marini et al. (12). In order to screen for curvature consistent with the wedge model of Trifonov and Sussman (13), we have considered the distribution of both a particular dinucleotide (e.g., ApA) as well as its complementary dinucleotide on the same strand (e.g., TpT). Assuming that a specific pair of adjacent bases curves the helix axis in one direction, one can expect curvature in the opposed direction to occur at the complementary dinucleotide. Thus the net curvature, within the context of the wedge model, depends on the distribution of both the 'wedge' dinucleotide and its complement along one strand.

TABLE I

DNA curvature scores and numbers of A and T tracts for the upstream region (-40 to -150)[a] of procaryotic promoters.

Promoter	Curvature Score	A_n tracts n>4	T_n tracts n>4	A_3	T_3
rrnD P1	12.3	6	2	1	0
rrnG P1	9.2	3	4	2	1
rrnH P1	5.65	5	1	1	0
rrnB P1	5.6	2	3	2	2
supBE	5.6	2	1	3	1
rrnC P1	5.55	3	3	2	1
lpp	5.3	3	2	0	1
hisR(Salm.)	2.8	2	2	3	1
his	2.6	3	1	2	1
rpoA	2.6	1	2	2	1
spoVG (B.sub.)[b]	2.6	1	2	2	1
spc	2.4	0	1	4	1
tnaA	2.2	2	1	2	5
uvrB P2	2.15	1	2	2	2
spot 42 RNA	2.1	2	2	1	0
tyrT	1.55	2	1	0	3
uvrB P1	1.55	0	3	2	1
deo P2	1.2	0	0	2	4
fol	1.2	1	2	0	3
rpoB	1.2	0	2	0	4
rrnA P2	1.2	1	0	3	0
rrnC P2	1.2	1	0	3	0
S10	1.2	2	0	2	3
rrnE P2	0.6	1	1	2	0
rrnG P2	0.6	1	0	3	0

Promoter	Curvature Score	A_n tracts n>4	T_n tracts n>4	A_3	T_3
rrnH P2	0.6	1	1	2	0
tufB	0.6	3	1	1	1
λP_L	0.6	0	2	0	4
ampC	0.	1	1	1	1
araBAD	0.	2	2	0	2
aroH	0.	1	0	0	0
hisA	0.	1	1	0	2
lac	0.	0	0	1	1
malEFG	0.	1	1	1	1
malK	0.	1	1	1	1
Pori-L	0.	0	0	1	2
Pori-R	0.	0	1	0	1
recA	0.	1	0	0	1
rplJ	0.	1	0	3	0
rrnD P2	0.	3	0	2	0
str	0.	0	0	1	2
trp P2	0.	2	0	1	1
trpR	0.	0	0	1	1

a - Region endpoints vary by a few nucleotides if the endpoints stated lie within an A or T stretch.
b - Score was evaluated for the region -40 to -80.

The autocorrelation function for a dinucleotide NpN within a sequence of length L is (12):

$$G_j = \sum_{i=1}^{L-j-1} \delta_i \, \delta_{i+j} \Big/ \sum_{i=1}^{L-j-1} \delta_i \qquad [1]$$

where $\delta_i = 1$ if an NpN starts at base pair i and zero otherwise. The positional correlation function for an NpN followed by its complement N'pN' j

TABLE II

DNA curvature scores and numbers of A and T tracts for the downstream region (+60 to +170) of procaryotic promoters.

Promoter	Curvature Score	A_n tracts n>4	T_n tracts n>4	A_3	T_3
S10	3.1	1	1	0	1
lac	1.55	1	1	0	1
malEFG	1.55	1	1	0	2
malK	1.2	1	0	2	1
recA	1.2	1	0	5	1
rrnB P1	1.2	2	0	2	2
rrnC P1	1.2	2	0	4	2
rrnA P2	1.2	1	0	1	2
rrnB P2	1.2	1	0	1	2
rrnC P2	1.2	1	0	1	2
rrnD P2	1.2	1	0	1	2
rrnE P2	1.2	1	0	1	2
rrnG P2	1.2	1	0	1	2
rrnH P2	1.2	1	0	1	2
rrnG P1	0.6	2	0	2	2
rrnH P1	0.6	1	0	4	2
spot 42 RNA	0.6	1	2	1	2
str	0.6	2	0	2	1
ampC	0.	1	0	1	1
aroH	0.	0	0	1	1
deo P2	0.	1	0	3	0
fol	0.	1	0	3	1
hisA	0.	0	0	0	1
his	0.	0	0	1	1
lpp	0.	1	0	1	0

Promoter	Curvature Score	A_n tracts n>4	T_n tracts n>4	A_3	T_3
Pori-R	0.	1	1	0	2
Pori-L	0.	1	2	1	1
rplJ	0.	1	0	0	1
rpoA	0.	0	0	1	1
rpoB	0.	2	0	1	1
rrnD P1	0.	0	0	4	2
supBE	0.	0	1	1	0
tnaA	0.	1	0	1	3
trp P2	0.	0	1	3	0
trpR	0.	1	1	0	1
tufB	0.	0	0	2	0

base pairs away is given by

$$F_j = \sum_{i=1}^{L-j-1} \delta_i^{\,1}\, \delta_{i+j}^{\,2} \Big/ \sum_{i=1}^{L-j-1} \delta_i^{\,1} \qquad [2]$$

where $\delta_i^{\,1} = 1$ if an NpN starts at base pair i, and $\delta_k^{\,2} = 1$ if an N'pN' starts at base pair k. $\delta_i^{\,1}$ and $\delta_k^{\,2}$ are zero otherwise. In order to determine if wedge model curvature is present within a region, the F_j plot for ApA followed by TpT was shifted by five base pairs and averaged with the G_j plot of ApA for $j \geq 7$. For $j \leq 6, G_j$ is plotted against j. A similar composite plot is determined by shifting the F_j plot for TpT followed by ApA and averaging it with the G_j plot of TpT for $j \geq 7$. We refer to these composite plots as the FG plots of ApA and TpT respectively.

RESULTS

Tables I and II list DNA curvature scores for forty-three promoter upstream regions and thirty-six downstream regions. The latter regions provided control sequences for the analysis. The major focus was on E. coli chromosomal promoters for which sequence data was available at least 150 bp upstream of the startpoint. A few other promoter regions were also examined. Most sequences were obtained from GenBank® Genetic Sequence Data Bank. The

first column of numbers shows the DNA curvature score resulting from the algorithm described in methods. Curvature scores were also calculated for the upstream regions omitting weights for the A_3 and T_3 tracts. The results were essentially the same. High transcription rate promoters are correlated with high DNA curvature scores. Four promoters which exhibit upstream activation - rrnB P1, tryT, hisR, and spoVG - have curvature scores above 1.5 between -40 and -150. The λP_L promoter which also shows upstream activation has a curvature score of 0.6 in this region. However if one examines the -65 to -175 region the score becomes 2.1. Deletions which reduce upstream activation reduce the curvature scores. For example, the hisR1223 mutation reduces its curvature score to 1.6. The rRNA P1 promoters have the highest scores of the promoters examined. These promoters can be responsible for 36-50% of the total transcription initiated in E. coli (14,15). The lpp promoter which initiates transcription for the major outer membrane lipoprotein in E. coli, is another very strong promoter, and it also has a high DNA curvature score. Several of the other promoters which have significant curvature scores (his, tyrT, spot42 RNA, hisR) also have high transcription initiation rates (3,16,17,18). Since the -35 and -10 regions of a promoter are clearly the major determinants of promoter strength, this correlation suggests that the upstream regions can modulate transcriptional activity. This is consistent with experimental studies on upstream regions (4). Table I also indicates that promoters which tend to have lower transcription initiation rates in vivo, such as the rRNA P2 promoters and the carbon source catabolic operon promoters, have low DNA curvature scores ($\leq$ 1.2). The number of A and T tracts over four base pairs long for each region, as well as the number of A_3 and T_3 tracts are given in Table I. Several regions have three or more tracts, but have a zero curvature score due to the lack of appropriate phasing of these tracts. None of the regions shows more than one A_3T_3 tract. Analysis of the +60 to +170 gene coding regions for thirty-six operons are given in Table II. Only three have curvature scores above 1.2. These are the lac P1, malEFG, and S10 downstream regions which have curvature scores of 1.55, 1.55, and 3.1, respectively.

The positional correlation analysis was also carried out on the regions described in Tables I and II. Composite FG plots were obtained using L = 110 bp. To determine a criteria for selecting regions which have curved helices in accordance with the wedge model, the upstream region of the E. coli rrnB P1 promoter was examined. The FG plot is shown in Figure 1. The four highest peaks in the ApA plot occur at j values of 10, 18, 27, and 37. Since DNA

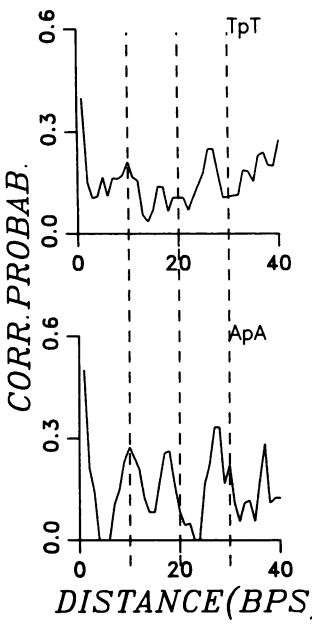

Figure 1. ApA and TpT FG plots for the -150 to -40 region of the E. coli rrnB P1 promoter.

fragments containing the rrnB P1 upstream region exhibit anomalous gel mobilities, other sequences were screened for a similar distribution of peaks. If either the ApA or TpT FG plots have their three highest peaks at j values of $9 \leq j \leq 11$, $18 \leq j \leq 22$, and $27 \leq j \leq 33$ the region is regarded as meeting this criteria. Only four out of the forty-three promoter upstream regions have ApA/TpT periodicities at helical repeat spacings. Examination of the thirty-six downstream sequences in Table II showed three sequences with ApA or TpT FG plots similar to figure 1. The ApA/TpT distributions which are periodic with the helix repeat do not correlate with promoter strength.

DISCUSSION

The results of this work show that a number of transcriptionally active promoters have upstream sequences which can be expected to produce curvature in the DNA helical axis. Promoters with known upstream activation regions have DNA curvature scores between 1.55 and 5.6 for a 110 bp stretch adjacent to the -35 site. The findings support the suggestion of Bossi and Smith (4) that DNA bending is a feature of the upstream activation of strong promoters. The data of Table I predicts that all of the rRNA P1 promoters have upstream regions which exhibit DNA curvature. Regions with high DNA curvature scores would be expected to exhibit anomalous mobility by the gel electrophoresis

assay. It will be of interest to examine fragments containing these regions for the structural feature of DNA bending, and the corresponding functional feature of upstream activation.

The DNA curvature score provides a simplified method of screening DNA regions of a given length for sequences which have a propensity to bend the helical axis. Although the magnitude of this scalar quantity should be correlated with the amount of curvature (i.e., the radius of curvature), a direct relationship can not be assumed. The relation between DNA sequence and helix curvature is not yet understood. Further studies on the effect of sequence on curvature will improve the weighting scheme we have employed. The extent to which stable helix curvature is involved in upstream activation is also not understood, although models can be proposed.

Since upstream activation can be demonstrated in vitro with RNA polymerase and linear DNA and no additional macromolecules, the mechanism may only involve these two components. Travers and coworkers have suggested that a secondary RNA polymerase binding site occurs in the upstream region (3,19). Bossi and Smith (4) have proposed that the promoter upstream region bends around the RNA polymerase increasing the number of specific polymerase-DNA contacts. One may combine these two ideas into two testable models. We consider these models within the context of the currently understood kinetic mechanism of RNA polymerase-promoter interactions.

Evidence has been given by several workers that RNA polymerase initially forms a specific 'closed' complex within a promoter site which converts to an 'open' complex through at least two sequential kinetic steps (20-22). One can describe this process as

$$R + P \underset{}{\overset{K_1}{\rightleftharpoons}} RP_c \overset{k_2}{\rightleftharpoons} I \rightleftharpoons RP_o \qquad [3]$$

where R and P designate the RNA polymerase and promoter site respectively, RP_c is the closed complex, RP_o is the open complex and I is an intermediate complex representing an isomerization of the RP_c complex. One model of how upstream regions may enhance transcription initiation is that they provide binding sites which increase the RNA polymerase concentration in the vicinity of the promoter. These stronger than average non-promoter sites could be due to the phased bends in the DNA, i.e., tertiary structure, rather than specific base pair sequences. If this model is valid, one could expect regions of DNA curvature to compete better for RNA polymerase than average quasi-random sequence DNA. Additionally, upstream activation should effect the initial

equilibrium binding step of forming RP_c, and not the formation of I.

A second model considers the possibility that the curvature of the DNA upstream region allows a segment of it to contact RNA polymerase after the polymerase has formed a closed complex at the promoter site. This step could be part of the isomerization to the intermediate complex I. This model provides several predictions. First, the effect of an upstream activation region would occur at the rate limiting isomerization step k_2. Second, an upstream region would, at least transiently, be in contact with RNA polymerase at a location differing from those interacting with the -35 and -10 regions of the promoter. Third, the direction of DNA bending would have to be specifically oriented to allow for this secondary DNA-polymerase contact. Curvature in the upstream region would be a necessary, but not sufficient condition for an activation effect. Neither of these models has substantive experimental support. Their presentation may be helpful in formulating tests of the mechanism of upstream activation.

ACKNOWLEDGEMENTS

We wish to thank the National Institutes of Health for support of this work through grant GM-33543 and Audrey Ralston for typing this manuscript. We also thank C. McCampbell for his helpful comments on the manuscript. Special thanks goes to R. L. Gourse for his helpful information.

REFERENCES

1. Hawley, D. K. and McClure, W. R. (1983) Nucl. Acids Res. 11, 2237-2255.
2. Horn, G. T. and Wells, R. D. (1981) J. Biol. Chem. 256, 2003-2009.
3. Lamond, A. I. and Travers, A. A. (1983) Nature 305, 248-250.
4. Bossi, L. and Smith, D. M. (1984) Cell 39, 643-652.
5. Gourse, R. L., de Boer, H. A., and Nomura, M. (1986) Cell 44, 197-205.
6. Banner, C. D. B., Moran, C. P. Jr., and Losick, R. (1983) J. Mol. Biol. 168, 351-365.
7. Hagerman, P. J. (1985) Biochemistry 24, 7033-7037.
8. Diekmann, S. (1986) FEBS Letters 195, 53-56.
9. Koo, H.-S., Wu, H.-M. and Crothers, D. M. (1986) Nature 320, 501-506.
10. Trifonov, E. N. (1985) CRC Crit. Rev. Biochem. 19, 89-106.
11. Hagerman, P. J. (1986) Nature 321, 449-450.
12. Marini, J. C., Levene, S. D., Crothers, D. M., and Englund, P. T. (1982) Proc. Natl. Acad. Sci. USA 79, 7664-7668.
13. Trifonov, E. N. and Sussman, J. L. (1980) Proc. Natl. Acad. Sci. USA 77, 3816-3820.
14. Lazzarini, R. A. and Dahlberg, A. E. (1971) J. Biol. Chem. 246, 420-429.
15. Pettijohn, D. E., Clarkson, K., Kossman, C. R., and Stonington, O. G. (1970) J. Mol. Biol. 52, 281-300.
16. Verde, P., Frunzio. R., di Nocera, P. P., Blasi, F., and Bruni, C. B. (1981) Nucl. Acids Res. 9, 2075-2086.
17. Sahagan, B. G. and Dahlberg, J. E. (1979) J. Mol. Biol. 131, 593-605.
18. Bossi, L. (1983) Mol. Gen. Genet. 192, 163-170.

19. Travers, A. A., Lamond, A. I., Mace, H. A. F., and Berman, M. L. (1983) Cell 35, 265-273.
20. Rosenberg, S. Kadesch, T. R., and Chamberlin, M. J. (1982) J. Mol. Biol. 155, 31-51.
21. Roe, J.-H., Burgess, R. R., and Record, M. J. Jr. (1984) J. Mol. Biol. 176, 495-521.
22. Buc, H. and McClure, W. R. (1985) Biochemistry 24, 2712-2723.